Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol 3, No.4, 2012

Segmentation and Feature Extraction of Tumors from Digital Mammograms

PRADEEP N *,

Department of Computer Science and Engineering, Bapuji Institute of Engineering and Technology, Davangere, Karnataka, INDIA.

Email: nmnpradeep@yahoo.com

GIRISHA H,

Department of Computer Science and Engineering, Rao Bahadur Y Mahabaleswarappa Engineering College, Bellary, Karnataka, INDIA.

KARIBASAPPA K,

Department of Computer Science and Engineering, Dayananda Sagar College of Engineering,

Bangalore, Karnataka, INDIA.

Abstract

Mammography is one of the available techniques for the early detection of masses or abnormalities which is related to breast cancer. Breast Cancer is the uncontrolled of cells in the breast region, which may affect the other parts of the body. The most common abnormalities that might indicate breast cancer are masses and calcifications. Masses appear in a mammogram as fine, granular clusters and also masses will not have sharp boundaries, so often difficult to identify in a raw mammogram. Digital Mammography is one of the best available technologies currently being used for the early detection of breast cancer. Computer Aided Detection System has to be developed for the detection of masses and calcifications in Digital Mammogram, which acts as a secondary tool for the radiologists for diagnosing the breast cancer. In this paper, we have proposed a secondary tool for the radiologists that help them in the segmentation and feature extraction process.

Keywords: Mammography, Breast Cancer, Masses, Calcification, Digital Mammography, Computer Aided Detection System, Segmentation, Feature Extraction

1. INTRODUCTION

Cancer is an abnormal, continual multiplying of cells. The cells divide uncontrollably and may grow into adjacent tissue or spread to distant parts of the body. Breast cancer remains a leading cause of cancer deaths among women in many parts of the world. Early detection of breast cancer through periodic screening has noticeably improved the outcome of the disease [1]. The mass of cancer cells will eventually become large enough to produce lumps, masses, or tumors that can be detected. Tumor is uncontrolled growth of cells which can be either Benign or Malignant. Benign Tumors are not cancerous. Benign tumors may grow larger but do not spread to other parts of the body. Malignant Tumors is cancerous. Malignant tumors can invade and destroy nearby tissue and spread to other parts of the body. Tumor can be easily identified in mammogram because tumor part is highly bright (having high intensity) compared to other part

(background) of the mammogram image as shown in Figure 1.

In the figure 1, we can observe that the marked oval shape area have higher intensity compared to the surrounding area. This marked oval shape is the required Region Of Interest (ROI). Segmentation techniques can be used to extract the tumor from the mammogram. Computer Aided (CA) detection systems have been developed to aid radiologists in detecting mammographic abnormalities [2-5].

There are large numbers of diagnostic methods currently available for the diagnosis of Breast Cancer, among which digital mammography is the most reliable method, for detecting early breast cancer [6-7]. Early detection of Breast Cancer can decrease the mortality rate. In this paper, we have proposed a CAD system to segment the tumor from the ROI and calculate the features that have importance in the classification of tumors.

2. PROPOSED CAD SYSTEM

The proposed CAD system consists of five principal stages. The first stage is the collection of input images and ROI extraction, second stage is ROI Image Enhancement, the third stage is Segmentation, the fourth stage is Filter and the last stage is Feature Enhancement.

2.1 ROI Extraction

The input images are collected from Mammographic Image Analysis Society (MIAS) research database. The MIAS dataset [6] is used to test the proposed technique. These images were previously investigated and labeled by an expert radiologist based on technical experience and biopsy. The dataset is selected due to the various cases it includes. The dataset is composed of 322 mammograms of right and left breast, from 161 patients, where 51 were diagnosed as malignant, 64 as benign and 207 as normal. The abnormalities are classified into microcalcification, circumscribed mass, ill-defined mass, spiculated mass, architectural distortion, and asymmetry. In this study 322 mammogram images were selected as described in Table 1. The original mammograms are 1024 x 1024 pixels, and almost 50% of the image comprised of the background with a lot of noise. Therefore a cropping operation is applied to the images to cut off the unwanted portions of the images. Regions of Interest (ROI) of 128x128 pixels are cropped manually, where the given center of the abnormality area is selected to be the center of ROI. In the proposed CAD system, microcalcifications are not considered, only circumscribed mass, ill-defined mass, spiculated mass, architectural distortion, and asymmetry are considered. The preprocessing phase of digital mammograms refers to the enhancement of mammograms intensity and contrast manipulation, noise reduction, background removal, edges sharpening, filtering, etc.

2.2 ROI Image Enhancement

Histogram equalization [7] has been used for the reassignment of pixels to make the image better, as proposed by H.D.Cheng et al [8]. In general, low contrast of mammographic images, hard to read masses in mammograms [9], variation of the intensities of the masses such that radiopaque mass with high density and radiolucent mass with low density in comparison with the background were found to be the reasons for the enhancement [10]. Selective median filtering with CLAHE (Contrast Limited Adaptive Histogram Equalization) algorithm will remove the noise and enhances the mammogram for better segmentation.

2.3 Segmentation

In analyzing mammogram image, it is important to distinguish the suspicious region from its surroundings. The goal of segmentation is to simplify and change the representation of an image into something that is more meaningful and easier to analyze. Image segmentation is typically used to locate objects and boundaries (lines, curves, etc.) in images. The result of image segmentation is a set of segments that

collectively cover the entire image, or a set of contours extracted from the image [11].

Segmentation can be carried out using any of the standard techniques like Local Thresholding, Global Thresholding, Region Growing, Region Clustering, Otsu Segmentation Technique, Template Matching. In the proposed system, we have used Local Thresholding Technique [12 -14] and Otsu method [15] for segmentation.

Local thresholding technique has been proven to provide an easy and convenient way to perform the segmentation on digital mammogram. The segmentation is determined by a single value known as the intensity threshold value. Then, each pixel in the image is compared with the threshold value. Pixel intensity values higher than the threshold will result in a white spot in the output image

The second technique used for segmentation is Otsu's method, which has shown a more satisfactory performance in the medical image segmentation. It has been found to perform well compared to other thresholding methods in segmenting the masses in digital mammogram.

2.4 Filter

After the segmentation process, there are still noises in the image which will affect the next stage, Feature Extraction stage. These noises have to be suppressed using any of the filtering techniques that are appropriate. Usually noises occupy less number of pixels compared to the tumor. These noises can be easily be visualized by observing the segmentation output and they look like small white spots. The observed pixels which are to be noises are removed by filling their intensity values by zeros.

2.5 Feature Extraction

Feature is used to denote a piece of information which is relevant for solving the computational task related to a certain application. More specifically, features can refer to:

- The result of a general neighborhood operation (feature extractor or feature detector) applied to the image,
- Specific structures in the image itself, ranging from simple structures such as points or edges to more complex structures such as objects.

Feature extraction is the important step in breast cancer detection. Texture feature is important for image classification. Various techniques have been used for computing texture features [16]

The recent use of textural features and machine learning (ML) classifiers has established a new research direction to detect breast cancer. Texture features have been widely used as classification of masses in digital mammogram. Texture features can be classified into three classes based on what they are derived from Gray Level Co-Occurrence Matrices (GLCM), Gray Level Difference Statistics (GLDS), and Run Length Statistics (RLS) [17]. Many studies have been focused on statistical texture features [18] used the three types of statistical texture feature for masses detection and classification, [19] used the RLS to extract the texture information from the region of interest. [20] Texture features in detection of masses in mammogram. [21] Used texture analysis for the classification of mammographic masses. [22] Used statistical texture features for early detection of masses in digitized mammogram.

Texture features is considered to be one of the widely used tool in masses detection in digital mammogram. The Haralick features and co-occurrence matrix are the most methods of textural analysis for feature extraction which are used to classify the regions of interest into benign and malignant masses. Most of the texture analysis methods are applied directly on the original or filtering images. The computational of the co-occurrence matrix to the whole image is required large computing time. However decreasing the size of the co-occurrence matrix by reducing the number of gray levels in the image without losing the information's is must.

Statistical texture features have been proven to be powerful in classifying masses and normal breast tissues

[23]. The implementation of feature extraction procedure relies on the quality of the texture, which is the main descriptor for all the mammograms.

In the proposed system, Co-occurrence matrix is used for feature extraction. GLCM is a powerful tool for image feature extraction. Gray level pixel distribution described by statistics like probability of two pixels having particular gray level at particular spatial relationships. This spatial information is provided as two dimensional gray level matrices. GLCM is a powerful tool for image feature extraction. Gray level pixel distribution described by statistics like probability of two pixels having particular gray level at particular spatial relationships. This spatial information is provided as two dimensional gray level matrices. Various statistical measures such as correlation, energy, entropy, homogeneity and other statistical measures are calculated based on GLCM for the output of fourth stage. The size of GLCM is determined by number of gray level in an image.

The standard Haralick GLCM texture descriptors from Haralick (1973), GLCM Texture Descriptors from Clausi (2002), GLCM Texture Descriptors from Soh and Tsatsoulis (1999) and MATLAB Image Processing Toolbox (IPT) GLCM Texture descriptors are depicted in table 2, table 3, table 4 and table 5 respectively. In the proposed system, combination of all the four texture descriptors and statistical features are extracted.

Apart from the texture features other statistical features like Mean, Standard Deviation, Variance, Skewness, Kurtosis, Root Mean Square is determined.

3. EXPERIMENTAL OUTPUT

MATLAB has been used to calculate all the features required and GUIDE toolbox is used to develop the GUI for the proposed system. In the first step, in the input image the ROI is selected by manual cropping process. For the cropped image, image enhancement techniques have been employed to improve the image quality. The segmentation is carried out for the enhanced cropped image by using Local thresholding and Otsu segmentation techniques. Before feature extraction of this segmented image, noise is removed by filling their intensity values by zero. These noises will appear as tiny white spots in the image. Features are calculated for the noise suppressed segmented output. The figure 3 depicts the steps carried out.

4. CONCLUSION AND FUTURE SCOPE

It has been observed that after the segmentation process, then also white spots-Noise will appear which will affects the feature extraction process. This noise has to be removed if feature extraction process has to yield the accurate results. Once features are calculated for the entire MIAS database, then feature set has to be constructed. The feature set has to be constructed in such a manner that it has to be understandable for the classifiers. Any of the classifiers like Artificial Neural Network, Support Vector Machines, and Decision Trees can be used for the classification of breast tumors. The accuracy of the classifier has to be determined. The accuracy of any classifier will depend on the features extraction phase. The classifier has to be trained for the feature set constructed. Then it has to be tested with unknown samples. The accuracy is calculated for training and testing phases. For the training and testing the classifiers, the MIAS dataset can be used. The classifiers can also be experimented for the digital mammograms collected from the local oncology hospitals.

References

[1] L. Tabar and P. B. Dean. Mammography and breast cancer: the new era. *Gynaecol Obstet*, 82:319–326, 2003.

[2] M. L. Giger, N. Karssemeijer and S. G. Armato, “Computer-aided diagnosis in medical imaging”, *IEEE*

Trans. on Med. Imaging, vol. 20, pp. 1205-1208, 2001.

[3] M. L. Giger, "Computer-aided diagnosis of breast lesions in medical images", *Comput. Science Engineering*, vol. 2, pp. 39-45, 2000.

[4] K. Doi, H. MacMahon, S. Katsuragawa, R. M. Nishikawa and Y. Jiang, "Computer-aided diagnosis in radiology: potential and pitfalls", *Eur. J. Radiol*., vol. 31, pp. 97-109, 1999.

[5] C. J. Vyborny, M. L. Giger and R. M. Nishikawa, "Computer-aided detection and diagnosis of breast cancer", *Radiologic Clinics of North America*, vol. 38, pp. 725-740, 2000

[6] http://peipa.essex.ac.uk/ipa/pix/mias.

[7] S.M.Pizer, E.O.P.Amburn and J.D.Austin, 1987. "Adaptive histogram equalization and its variations", *Comput. Vision Graphics Image Process. 39*, pp. 355-368.

[8] H.D.Cheng, X.J.Shi, R.Min, L.M.Hu, X.P.Cai and H.N.Du, 2006. "Approaches for Automated Detection and Classification of Masses in Mammograms", *Pattern Recognition 39*, pp. 646- 668.

[9] Jacob Scharcanski and Claudio Rosito Jung, 2006. "Denoising and enhancing digital mammographic images for visual screening", *Computerized Medical Imaging and Graphics 30*, pp. 243–254.

[10] R.M. Rangayyan, L. Shen, Y. Shen, J.E.L. Desautels, H. Bryant, T.J. Terry, N. Horeczko and M.S. Rose, 1997. "Improvement of sensitivity of breast cancer diagnosis with adaptive neighbourhood contrast enhancement of mammograms", *IEEE Trans. Inform. Technol.Biomed. 1*, pp. 161-170.

[11] G.M. Brake and N. Karssemeijer, 2001. "Segmentation of suspicious densities", *Med. Phys. 28*, pp. 258-266.

[12] N. Karssemeijer. Automated classification of parenchymal patterns in mammograms. *Phys. Med. Biol.*, 43:365–378, 1998.

[13] Sung-Nien Yu, Kuan-Yuei Li and Yu-Kun Haung, 2006. "Detection of microcalcifications in digital mammograms using wavelet filter and Markov random field model", *Computerized Medical Imaging and Graphics 30*, pp. 163-173.

[14] Xu Kai, Qin Kun and Pei Tao, 2006. "Interactive Method for Image Segmentation Based on Cloud Model", *Computer Engineering and Application 34*, pp. 33-35.

[15] Jeong HJ, Kim TY, Hwang HG et al. Comparison of thresholding methods for breast tumor cell segmentation. IEEE Proceedings of 7th International Workshop on Enterprise Networking and Computing in Healthcare Industry. 2005: 392-5.

[16] H.S.Sheshadri and A.Kandaswamy,"Experimental investigation on breast tissue classification based on statistical feature extraction of mammograms", Computerized Medical Imaging and Graphics, no.31, pp.46-48, 2007.

[17] H.D. Cheng, X.J. Shi, R. Min, L.M. Hu, X.P. Cai, H.N. Du (2006) "Approaches for automated detection and classification of masses in mammograms", Pattern Recognition, Vol. 39, pp. 646-668.

[18] Mavroforakis,M.E.,Georgiou,H.V.,Dimitropoulos,N., Cavouras,D., and Theodoridis,S., "Mammographic masses characterization based on localized texture and dataset fractal analysis using linear, neural and support vector

machine classifiers", *Artificial Intelligence in Medicine*, 2006, pp. 145—162.

[19]Szekely, N, Toth, N. Pataki, B. (2004). A hybrid system for detecting masses in mammographic images. Instrumentation and measurement technology conference, 2004. IMTC 04. Proceeding of the 21st IEEE Vol3, 18-20 May 2004 pp2065-2070.

[20] K. Bovis and S. Singh. Detection of masses in mammograms using texture features. 15th International

Conference on Pattern Recognition (ICPR'00), 2:2267, 2000.

[21] Mudigonda, N.R, Rangayyan, R. Desautels, J.E.L (2000); Gradients and texture analysis for the classification of mammographic masses Medical Image, IEEE Transaction on Vol 19, Issue 10, Oct. 2000

pp 1032-1043.

[22] N. Youssry, F.E.Z. Abou-Chadi, and A.M. El-Sayad. Early detection of masses in digitized mammograms using texture features and neuro-fuzzy model. 4th Annual IEEE Conf on Information Technology Applications in

Biomedicine, 2003.

[23] Haralick,R.M., Shanmugam,K., Dinstein,I., "Textural features for image classification", *IEEE Trans Sys ManCyb*, 1973, pp. 610—21.

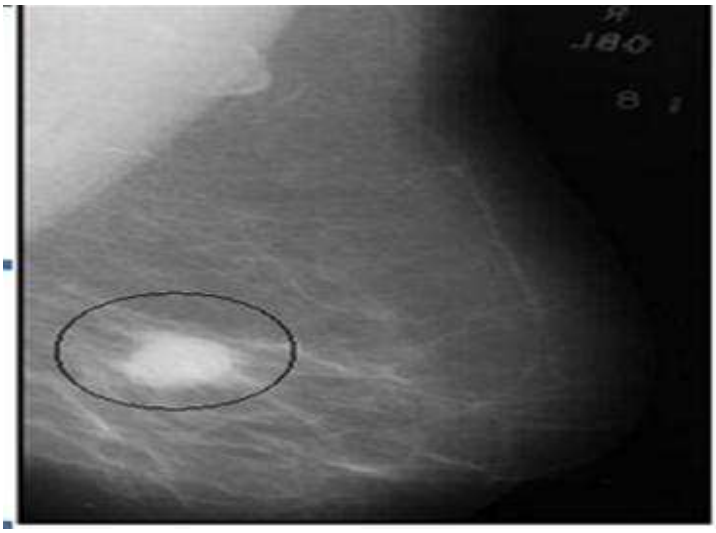

Figure 1. Sample mammogram

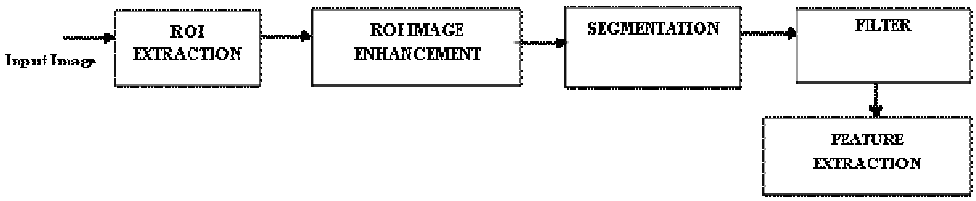

Figure 2. Proposed CAD System

Class	Benign	Malignant	Total
Microcalcification	12	13	25
Circumscribed	19	4	23
ll-defined	7	7	14
Spiculated	11	8	19
Architectural	9	10	19
Asymmetry	6	9	15
Normal tissue	9	-	207
Total	64	51	322

Table 1. The Distribution of MIAS Data Set

No.	Texture Descriptor	Formula
1.	Angular Second Moment: Energy	$\sum_i \sum_j \{p(i,j)\}^2$
2.	Contrast	$\sum_{n=0}^{N_g-1} n^2 \left\{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} p(i,j), \lvert i-j \rvert = n\right\}$
3.	Correlation (Haralick)	$\sum_i \sum_j \frac{(ij)p(i,j)-\mu_x\mu_y}{\sigma_x\sigma_y}$ where, μ_x, μ_y, σ_x and σ_y are the means and standard deviations of p_x and p_y, the partial probability density functions.
4.	Sum of Squares: Variance	$\sum_i \sum_j (i-\mu)^2\, p(i,j)$
5.	Inverse Difference Moment	$\sum_i \sum_j \frac{p(i,j)}{1+(i-j)^2}$
6.	Sum Average	$\sum_{i=2}^{2N_g} i p_{x+y}(i)$ where, x and y are the coordinates (row and column) of any entry in the co-occurrence matrix, and $p_{x+y}(i)$ is the probability of co-occurrence matrix coordinates summing to $x+y$.
7.	Sum Variance	$\sum_{i=2}^{2N_g} (i-S_{ent})^2\, p_{x+y}(i)$
8.	Sum Entropy	$-\sum_{i=2}^{2N_g} p_{x+y}(i) \log\{p_{x+y}(i)\} = S_{ent}$
9.	Entropy	$-\sum_i \sum_j p(i,j) \log(p(i,j))$
10.	Difference Variance	$\sum_{i=0}^{N_g-1} i^2 p_{x-y}(i)$
11.	Difference Entropy	$-\sum_{i=0}^{N_g-1} p_{x-y}(i) \log\{p_{x-y}(i)\}$
12.	Information Measure of Correlation 1	$\frac{HXY-HXY1}{\max\{HX,HY\}}$ where, HX and HY are the entropies of p_x and p_y such that: $HXY = -\sum_i \sum_j p(i,j) \log(p(i,j))$ $HXY1 = -\sum_i \sum_j p(i,j) \log\{p_x(i)p_y(j)\}$ $HXY2 = -\sum_i \sum_j p_x(i)p_y(j) \log\{p_x(i)p_y(j)\}$
13.	Information Measure of Correlation 2	$(1-\exp[-2(HXY2-HXY)])^{1/2}$
14.	Maximum Correlation Coefficient	$\sqrt{(\text{Second largest eigenvalue of } Q)}$ where, $Q(i,j) = \sum_k \frac{p(i,k)p(j,k)}{p_x(i)p_y(k)}$

Table 2: Standard GLCM Texture Descriptors (Haralick, 1973)

No.	Texture Descriptor	Formula
1.	Inverse Difference Normalized	$\sum \frac{c(i,j)}{1+\lvert i-j\rvert}$ where $c(i,j)$ is the co-occurrence probability between grey levels (i and j) defined as: $c(i,j) = \frac{p(i,j)}{\sum_{i,j=1}^{G} p(i,j)}$
2.	Inverse Difference Moment Normalized	$\sum \frac{c_{ij}}{1+(i-j)^2}$

Table 3: GLCM Texture Descriptors from Clausi (2002)

No.	Texture Descriptor	Formula
1.	Autocorrelation	$\sum_i \sum_j (ij) p(i,j)$ where $p(i,j)$ represents the number of occurrences of grey levels (i and j).
2.	Cluster Prominence	$\sum_i \sum_j (i + j - \mu_x - \mu_y)^4 p(i,j)$ where $p(i,j)$ is the (i,j)th entry in a normalized GLCM. The mean for the rows and columns of the matrix are: $\mu_x = \sum_i \sum_j i \cdot p(i,j)$ $\mu_y = \sum_i \sum_j j \cdot p(i,j)$
3.	Cluster Shade	$\sum_i \sum_j (i + j - \mu_x - \mu_y)^3 p(i,j)$
4.	Dissimilarity	$\sum_i \sum_j \lvert i - j\rvert \cdot p(i,j)$
5.	Homogeneity (Soh & Tsatsoulis)	$\sum_i \sum_j \frac{1}{1+(i-j)^2} p(i,j)$
6.	Maximum Probability	$\max_{i,j} p(i,j)$

Table 4: GLCM Texture Descriptors from Soh and Tsatsoulis (1999)

No.	Texture Descriptor	Formula
1.	Correlation (MATLAB)	$\sum_{i,j} \frac{(i-\mu_x)(j-\mu_y)p(i,j)}{\sigma_x \sigma_y}$ where $p(i,j)$ is the (i,j)th entry in a normalized GLCM. The standard deviations for the rows and columns of the matrix are: $\sigma_x = \sum_i \sum_j (i-\mu_x)^2 \cdot p(i,j)$ $\sigma_x = \sum_i \sum_j (j-\mu_y)^2 \cdot p(i,j)$ where μ_x and μ_y are the means for the rows and columns of the matrix respectively
2.	Homogeneity (MATLAB)	$\sum_{i,j} \frac{p(i,j)}{1+\lvert i-j \rvert}$

Table 5: GLCM Texture Descriptors from the MATLAB Image Processing Toolbox

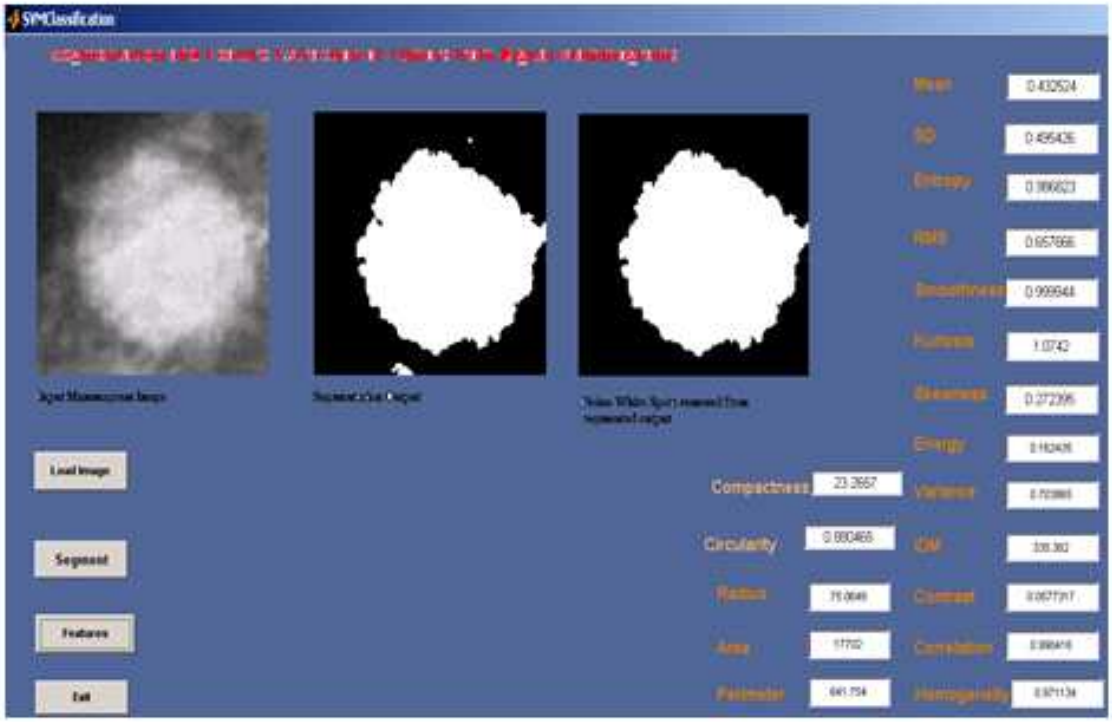

Figure 3. Experimental Output Snapshot